A DEDUCTIVE ENQUIRY SYSTEM FOR A PALAEONTOLOGICAL DATABASE OF MUSEUM MATERIAL

by M. J. ROGERS, D. T. DONOVAN *and* M. H. ROGERS

ABSTRACT. The use of database management systems for cataloguing museum material is becoming widespread. Recent advances in database management systems indicate that in future such systems will have a range of increased facilities, including deductive capabilities and natural language interfaces for casual enquiries. In order to explore some of the consequences of these developments for museum catalogues, a prototype system has been written and examples of its use demonstrated. These show that such systems will have significant research potential in a number of areas in geology in addition to taxonomy.

OUR museums contain a wealth of information which can be used in taxonomic, stratigraphic, palaeoecological and other geological studies. If these valuable resources are to be fully exploited, a sophisticated catalogue system is necessary.

The first catalogues of fossils, or registers of specimens, were usually impressive books containing a list of specimens numbered according to some local system, together with information concerning geographic, stratigraphic and collection details. The entries were made in beautiful copperplate writing but there was little attempt at ordering the entries. To obtain information from such registers is almost impossible unless the registration number of the required material is known. The register might be supplemented by card-index systems, one based on genus, one on stratigraphic horizon, one of type and figured specimens and so on. Such indexes increase the usefulness of museum material, but they are onerous to produce and bulky to store.

During the past twenty years progress has been made towards storing museum catalogue information on computer databases. Price (1984) described how a hierarchical database management system was designed by Cutbill *et al.* (1971) for material in the Sedgwick Museum, Cambridge. Several museums now use improved versions of this early system, such as GOS (Porter 1982, 1983), which was released by the Museums Documentation Association, or a more sophisticated version MUSCAT. Other museums, including Bristol University Museum, use commercially available database management systems, for example the relational databases dBASE III or Oracle. International progress in this field was reviewed by Roberts and Light (1980), and Brunton *et al.* (1985) discuss computing procedures in their *Guidelines for the curation of geological material.*

These systems have enormously increased the availability of information stored in our museums because explicit queries can be made on any attribute recorded in the database, and lists are easily prepared in which specimens are ordered alphamerically on a particular attribute such as name, collector, geographic location or stratigraphic horizon.

Research into database management systems has continued at a high level, and during the past few years advances have been made in several areas including database query languages, user interfaces, deductive databases and recursive query processing. At the same time there has been an increase in the use of remote databases and distributed databases, and many of these developments will be incorporated into the next generation of commercial systems. Much of this work has been done using logic programming languages, in particular Prolog, which have greatly facilitated the construction of prototype systems incorporating some or all of these developments. Thus it is possible to explore the consequences of these advances before commercial systems become widely

[Palaeontology, Vol. 33, Part 3, 1990, pp. 613–622.]

available. The relevance of some of these developments has already been recognised in a wide range of disciplines; Rawlings (1988) gives a discussion of applications to molecular biology. Here we investigate the consequences of applying these concepts to the cataloguing of museum material. We describe our experience in designing a prototype system for the large Jurassic ammonites in the University Museum, Bristol, and give some illustrations of the power of such a system in other fields of geology, not just as a museum tool.

PROLOG AND DATABASES

In this section we give a brief explanation of why Prolog is particularly suitable as a research tool in this area; for a full description of the language the reader is referred to one of the many excellent books on the subject, for example Sterling and Shapiro (1986). The recent book edited by Gray and Lucas (1988) gives a good account of recent developments in the use of Prolog in conjunction with databases.

Prolog (programming in logic) is a programming language based on logic and was developed by Colmerauer and his colleagues at Marseille in 1970 (Colmerauer *et al.* 1973). The use of the language has grown rapidly, especially when it became clear in 1981 that logic programming would play a fundamental role in the Japanese Fifth Generation Project. Prolog is now used in a wide variety of applications including expert systems, artificial intelligence, natural language processing, deductive databases and general problem solvers. An example in palaeontology is described by Brough and Alexander (1986) who use Prolog to construct an expert system which provides computer assisted identification of fossils. It has become clear that there is a close relationship between Prolog and relational databases (Gallaire *et al.* 1984), and more generally between Prolog and database query languages. As an example the Prolog predicate

```
f(X,Z):-g(X,Y),h(Y,Z)
```

has the declarative meaning that f(X, Z) is true for particular values of X and Z if g is true for X and some value Y, and h is true for Y and Z. But in a relational database language this clause has the interpretation that the relation f is the join of relations g and h. Nevertheless, there are major differences between these two approaches. For example, Prolog possesses great flexibility in the data structures it can handle, and allows recursion in both rules and facts. Relational database management systems, on the other hand, have advantages in such areas as updating, efficient handling of transactions, and concurrency.

There is currently a considerable amount of interest in combining these facilities in order to produce a new generation of database query systems. One approach is to use Prolog to construct a 'front end' which transforms a query into a standard form for execution on an existing database management system (DBMS), and an example of this is given in Ghosh *et al.* (1988). Clearly, developments along these lines which allow queries to be formulated in restricted natural language will be of great benefit for casual enquiries to existing databases. Another approach has been to establish deductive database systems containing predicates (or rules) which can be used to produce facts (or tuples) which are not stored explicitly in the database.

In the present work we are examining a prototype system written in Prolog, where the data have come partly from an existing relational database system containing the University of Bristol collection of large ammonites, while the remaining data, the knowledge base, consisting of appropriate chronostratigraphic and ammonite classification schemes and other relevant data, have been entered *ab initio*. The aim is to illustrate, by means of a few examples, the type of query which can be handled and to indicate some of its potential uses. No attempt has been made in this work to produce a natural language front end; as explained above there is intense commercial activity in this area and such products will undoubtedly become widely available in the future.

THE KNOWLEDGE BASE: PROLOG DATA STRUCTURES FOR GEOLOGICAL DATA

General

The particular aspects of geology which are discussed here are chronostratigraphy, palaeontological classification, lithological horizon and locality. These aspects are of two main types. Chronostratigraphic and palaeontologic classifications are both attempts to impose order on imperfectly understood phenomena and are of a hierarchical form. They can be represented by a tree structure in Prolog, but the quadtree, a particular form of tree used in geographic information systems (Williams 1988), is too inflexible for the present purpose. The data structures defined in this paper are more general as is necessary to accommodate the complexities of a chronostratigraphic and palaeontologic information system. Both classifications have been modified over the years – and will no doubt continue to be modified in the light of future research. In consequence, both types of classification contain many obsolete terms. In contrast, lithological horizon and locality are observed facts. Again, the same stratum or the same outcrop may be known by more than one name, but in each case a name is associated with an observation, not an abstraction.

In the case of the classification schemes, not only is it necessary to define the structures, but also the data must be internally consistent. As noted above, these schemes are continuously evolving. However, in order to maintain internal consistency, it is necessary to choose a scheme and adhere to it. The choice of a classification for a particular taxon or stratigraphic system is a task for a specialist who should ensure that it is modern, widely accepted and readily available in the literature. The Special Reports published by the Geological Society could be an appropriate source for British chronostratigraphy and the Jurassic volumes (Cope, Getty *et al.* 1980; Cope, Duff *et al.* 1980) are selected here for the Jurassic system. The recent revision of the *Treatise on invertebrate paleontology* is used for the ammonite classification (Donovan *et al.* 1981).

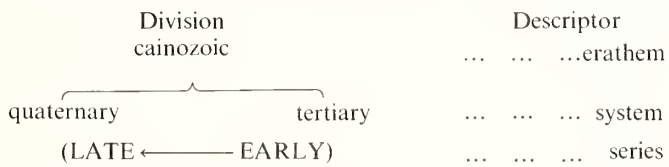
Because the lithological horizon and location are factual records, there is not the same problem of internal consistency. In this prototype only one set of lithostratigraphic terms is used at each locality, and no attempt has been made to incorporate a hierarchical classification for lithostratigraphy.

Data structures for chronostratigraphic classification

The chronostratigraphic classification is a tree structure with two additional properties:

1. The sub-trees (which can contain an arbitrary number of elements) are ordered in a time sense, where the latest appears first in the list and the earliest (oldest) last.
2. Each level of the tree is referred to by a descriptor, which is used by geologists as a chronostratigraphic division.

Thus, for example, a small subset of the scheme can be represented as:



At the system level, Quaternary is later than Tertiary.

The actual representation used in this report has the following chronostratigraphic divisions (the descriptors):

erathem – system – series – stage – substage – zone – subzone

Each division is not always used. For example, the Aalenian Stage is divided into zones, whereas the Bajocian Stage has two substages, the Upper Bajocian and Lower Bajocian, both of which are divided into zones. The Prolog representation has been chosen so that it can accommodate such

variations of the basic data structure, and accordingly contains three arguments. The first is the name of the node and will be a particular system or series etc. The second is the name (descriptor) of its subdivisions, and the third is a list of those subdivisions with the convention that the elements in the list are ordered, with the latest first and the earliest last. Thus, the general definition is of the form

```
tree(Node,Subdivision,List)
```

and the particular example illustrated above is written as

```
tree(cainozoic,system,[quaternary,tertiary]).
```

A chronostratigraphic scheme based on this representation has been written down to the level of subzone divisions for the Jurassic, using data given in Cope, Getty *et al.* (1980) and Cope, Duff *et al.* (1980).

A further complication in chronostratigraphy is that obsolete terms are met when studying specimens from old collections, or in older literature. The following simple structure is used to incorporate such terms.

```
obsolete(Name,Division,Youngest_division,Oldest_division)
```

where Name is the obsolete name, Division is the chronostratigraphic level in the structure 'tree' of the terms Youngest_division and Oldest_division which are the upper and lower boundaries of Name according to Cope, Getty *et al.* (1980) and Cope, Duff *et al.* (1980). For example, the Prolog fact defining the Charmouthian is written as

```
obsolete(charmouthian,zone,davoei,jamesoni)
```

Using this representation we can now formulate in Prolog several rules of varying complexity which can be used in a query system. Some examples are given below.

1. stratum(Y,X).

This states that Y is a division of X so that for example stratum(quaternary,cainozoic) means that Quaternary is a division of the Cainozoic. The Prolog rule for this is:

```
stratum(Y,X):- tree(X,_L),member(Y,L).
```

The query ?-stratum(Y,cainozoic). will, if asked repeatedly, generate all the divisions of the Cainozoic, while the query ?-stratum(tertiary,X). will produce the single answer

```
X = cainozoic
```

2. substratum(Y,X).

This is similar to stratum(Y,X) defined above except that it spans an arbitrary number of 'generations'. The Prolog definition for this involves two rules:

```
substratum(Y,X):- stratum(Y,X).
substratum(Y,X):- stratum(Y,Z), substratum(Z,X).
```

The query ?-substratum(Y,cainozoic). will, if asked repeatedly, generate all the divisions down to, and including, the level of subzones which are in the Cainozoic Erathem. These rules are useful building blocks to which other rules can be added to formulate powerful queries. For example

```
?-list_x_to_y(X,Y,Lr).
```

lists all the elements, including X and Y, between X and Y, irrespective of the division to which they belong, or their order. Thus

```
?-list_x_to_y(obtusum, bucklandi,L),list(L).
```

will yield the following result

obtusum
turneri
semicostatum
bucklandi

even though the *obtusum* Zone is in the Upper Sinemurian and the remaining zones are in the Lower Sinemurian.

Data structures and Prolog representation for palaeontological classification

The palaeontological classification is a tree structure which is similar to that adopted for the chronostratigraphic scheme. The main difference is that the subtrees at each stage are unordered, whereas in the chronostratigraphic case there is a temporal ordering, with the latest first and the earliest last in each subtree. Each level of the tree is referred to by a descriptor, for example order, family, genus etc. A small example of the scheme is given by:

Division						Descriptor
arietitidae		...	...	...	...	family
arietitinae	agassiceratinae	...	...	...	...	sub-family
coroniceras	arnioceras	...	...	...	...	... genus

The hierarchical ordering of the levels or taxa is:

suborder – superfamily – family – sub-family – genus – subgenus

Some of these levels are absent in particular cases, so that, for example, a family could be composed of a set of genera. The general Prolog representation is of the form

```
ammclass(Node,Subdivision,List)
```

where Node is the name of a particular palaeontological category and Subdivision is the classificatory level of the items in the list. The small example above is accordingly represented as

```
ammclass(arietitidae,subfamily,[arietitinae,agassiceratinae ...]).
```

Using this representation, a query pedigree(X) has been written which will list all the ancestors of the argument X. The Prolog definition of pedigree is

```
pedigree(ammonitina):-!
pedigree(X):-division(X,D),write(X),tab(3),write(D),nl,stratum(X,Y),pedigree(Y).
```

and the query

```
?-pedigree(arietites).
```

yields the result

arietites	subgenus
coroniceras	genus
arietitinae	subfamily
psiloceratacae	superfamily

An additional complication in the palaeontological classification scheme arises through the

widespread occurrence of synonyms. To deal with this, a further Prolog representation has been introduced of the form

```
synonym(Name,Level,List_of_synonyms)
```

so that for example

```
synonym(schlotheimia,genus,[anguliferites,scamnoceras]).
```

indicates that *Anguliferites* and *Scamnoceras* are all synonyms of the genus *Schlotheimia*. This structure is used when dealing with specimens which were identified according to earlier classifications.

Link between the chronostratigraphic and palaeontologic classification

The two schemes described are independent, and indeed each is independent of any classification scheme. However, the link between the two is of paramount importance and accordingly a further Prolog structure has been designed to represent this link. This is of the form

```
pal_range(Name,Taxon,Division,Upper_occurrence,Lower_occurrence)
```

As an example of this

```
pal_range(caloceras,genus,zone,liasicus,planorbis).
```

represents the fact that the genus *Caloceras* occurs in the zone range *liasicus* to *planorbis*. It should be noted here that it is only necessary for entries in this category to be made at the generic or subgeneric level; the range for higher categories can easily be derived using simple Prolog constructs.

A query gen_range has been written which, by searching both the chronostratigraphic and classificatory trees, lists all genera found in the chronostratigraphic range R1 to R2. The program is as follows

```
gen_range(R1,R2):- list_x_to_y(R1,R2,L),!,pal_range(N,genus,_,X,Y),
    member(X,L),member(Y,L),
    nl,tab(15),write(N),fail.
```

so that for example the query

```
?-gen_range(obtusum,bucklandi).
```

gives the result

```
sulciferites
coroniceras
arnioceras
metarnioceras
tmaegoceras
agassiceras
euagassiceras
asteroceras
aegasteroceras
caenisites
epophioceras
xipheroceras
microderoceras
promicroceras
```

Prolog representation for observed data

The observed data of lithological horizon and location are intimately related, so the two types of information are combined in one data structure.

The general form is

```
placc(Grid_easting,Grid_northing,Locality_name,List)
```

List elements are of the form

```
correlation(Lithologic_horizon,Thickness,Chronostratigraphic_division,Upper_occurrence,
Lower_occurrence)
```

The fact

```
place(368300,165000,corston_field_quarry,[correlation(blue_lias,2·4,zone,angulata,angulata)]).
```

records the fact that the outcrop at Corston Field Quarry (368300,165000), consists of 2·4 metres of Blue Lias, all in the *angulata* Zone.

In this Prolog structure, the observed data of the lithological horizons recorded at a section is linked to the chronostratigraphy according to the current interpretation. A revision of this correlation in the light of new evidence can easily be incorporated by a simple modification of the data relating to that section.

Many queries can be asked using the data structures so far described which are of use to geologists in many fields and not only the taxonomist or museum curator.

One such is a query which lists the locations and ranges of a lithostratigraphic unit and its thickness in metres in the overall chronostratigraphic range R1 to R2.

```
?-thick(obtusum,bucklandi,blue_lias).
```

yields

```
saltford_midland_railway_cutting 3 semicostatum bucklandi
kelston_park_quarry 3·93 semicostatum bucklandi
```

This information is essential for the construction of palaeogeographic maps, geologic cross-sections and other visual representations of stratigraphy, and indeed the answers could be transferred directly as data to an appropriate graphics program.

THE DATABASE: DATA STRUCTURE FOR MUSEUM SPECIMENS

The data recorded about each specimen in a catalogue should be a list of facts including horizon, locality and collector and also the subsequent history of the material. A specimen may be moved from one collector or collection to another and it may be identified and cited or figured by several different workers at different times. Each of these events should be noted because they are often important in identifying type material.

In practice it is rare for all these facts to be known about a specimen, particularly if it was collected many years ago and has been moved from collection to collection, but it is often possible to make inferences about its origin and history. For example, if the handwriting on a specimen is recognised as that of a well known collector, it may be assumed that it was collected by that person. These inferences should be clearly noted in the catalogue.

In the prototype described here we make no attempt to create structures to accommodate all the complexities which should be recorded in a museum catalogue. Instead we have selected a few facts, which we assume to be correct, about each specimen so that we can indicate how the museum data can be used in conjunction with the knowledge base. Initially, these facts formed the columns in a

table of a relational database. This information has been transferred into a file and edited into a series of Prolog facts of the form

```
ammonite(V1,V2,V3,V4,V5,V6,V7,V8,V9,V10,V11,V12,V13)
```

where V1 and V2 are identifying numbers, V3,V4 are the generic and specific name, V5 = number of specimens, V6 = matrix, V7 = biological class or order, V8 = size in cm, V9 = stratigraphic system, V10 = horizon, V11 = locality, V12 = geographic region, and V13 = quality of specimen.

The relational database was designed and the ammonite data recorded before the current research was started. No attempt has been made so far to test its Prolog representation for internal consistency, or to ensure that all the terms in the ammonite file are present in the knowledge base. However, with very little editing, it has been possible to search for all specimens belonging to a particular taxon. In this query, the knowledge base is searched and all the members of the same taxon and their synonyms are listed, followed by all such specimens recorded in the database.

In the example, the genus *Schlotheimia* belongs to the family Schlotheimiidae and all the genera in this family and their synonyms are listed. Thus

```
?-taxon(schlotheimia).
```

yields

```
schlotheimia
synonyms are:
   anguliferites
   scamnoceras
angulaticeras
synonyms are:
   argoceras
   boucaulticeras
etc.
{the museum specimens belonging to the family Schlotheimiidae are:}
   28 5001 schlotheimia
    9 5001 schlotheimia
etc.
```

The answer to such a query is invaluable to the taxonomist. Further queries should be constructed which use the horizon and location data from the ammonite structure in conjunction with the knowledge base. The palaeobiologist could easily use these results to test hypotheses which would be very onerous to test by conventional means.

It is pertinent to ask whether the information contained in a computerized museum catalogue is of a quality suitable for research purposes. When using any secondhand material, whether from a museum collection or from published information, the researcher always has to balance the accuracy of the information about a specimen against the availability of alternative specimens and the time and expense of collecting and investigating fresh material.

In a museum catalogue the identification may be wrong, the identifier unknown and the locality and the horizon may have been inferred because of the lithology or some other characteristic feature of the specimen. The computer catalogue should contain all available evidence to enable a researcher to assess the quality of the information about a specimen. Also, if necessary, the specimen itself can be examined.

In the case of published information, it is not always possible to find the source material, particularly referenced material, and in such circumstances future researchers depend entirely upon the professional skill and integrity of the author(s).

SUGGESTIONS FOR FURTHER DEVELOPMENTS

The deductive database described in this paper could form the basis of a powerful information retrieval/on-line query system, but it will need to be extended and developed in several main areas. The knowledge base should be extended to include, for example, all stratigraphic systems and biological phyla. Facilities should be provided so that stratigraphic classifications appropriate to other parts of the world can be incorporated. The structures described in this paper are sufficiently general to achieve this, but the linkage between data from different provinces must be designed with care so that new evidence on correlation can easily be included.

The source of lithostratigraphic information should be added to the description of sections and research, and collection details should be included in the database of museum specimens.

A data entry facility should be provided which checks all new data for consistency and accuracy, as far as this is possible. A 'user-friendly' interface is essential for the casual user and this can be achieved either by a menu-driven query system or a limited natural language processor. The full power of the system described here can only be realized if it is integrated with geologically orientated program packages.

SUMMARY AND CONCLUSIONS

The primary aim of this paper is to show that the valuable information contained in geological museum catalogues can be made readily available for a wide range of investigations. This can be achieved by constructing a knowledge base of geological information and using it in conjunction with a conventional database of museum specimens. The knowledge base itself is an important reference source for many geological investigations. The information in the knowledge base is held in a series of simple Prolog structures which are linked at only a few essential points. Consequently, any modifications which may arise through the availability of new evidence or new interpretations can be made easily, and additional facts can be added without corrupting the existing information.

The geological information included in the conventional museum database can be restricted to observed facts, while the presence of checking routines, although not included in this prototype, would ensure that the internal consistency and accuracy of the database is maintained. This is particularly important when data are entered by relatively unskilled personnel and allows the specialist to concentrate on the complex problems.

The deductive capabilities which are available in the knowledge base allow a more general class of queries to be made than is possible with a conventional database management system.

The results presented in this paper are based on the use of local databases, but it is already clear from the experience currently being gained with remote access to library catalogues that the techniques described here can be extended to include remote databases, both at national and international sites.

REFERENCES

BROUGH, D. R. and ALEXANDER, I. F. 1986. The fossil expert system. *Expert Systems*, **3**, 76–83.

BRUNTON, C. H. C., BESTERMAN, T. P. and COOPER, J. A. 1985. Guidelines for the curation of geological materials. *Miscellaneous Papers of the Geological Society*, **17**, 1–174.

COLMERAUER, A., KANOUI, H., PASERO, R. and ROUSSEL, P. 1973. Un système de communication homme-machine en français. *Rapport de recherche sur le contract CRI*, **72–81**. Groupe d'Intelligence Artificielle, Université d'Aix-Marseille II, France.

COPE, J. C. W., GETTY, T. A., HOWARTH, M. K., MORTON, N. and TORRENS, H. S. 1980. A correlation of Jurassic rocks in the British Isles. Part 1. Introduction and Lower Jurassic. *Geological Society of London. Special Report*, **14**, 1–73.

COPE, J. C. W., DUFF, K. L., PARSONS, C. F., TORRENS, H. S., WIMBLEDON, W. A. and WRIGHT, J. K. 1980. A correlation of Jurassic rocks in the British Isles. Part 2. Middle and Upper Jurassic. *Geological Society of London. Special Report*, **15**, 1–109.

CUTBILL, J. L., HALLEN, A. J. and LEWIS, G. D. 1971. A format for the machine exchange of museum data. 311–320. *In* CUTBILL, J. L. (ed.). *Data processing in biology and geology*, Academic Press, London, 346 pp.

DONOVAN, D. T., CALLOMON, J. H. and HOWARTH, M. K. 1981. Classification of the Jurassic Ammonitina. 101–155. *In* HOUSE, M. R. and SENIOR, J. R. (eds.). *The Ammonoidea*, Academic Press, London, 593 pp.

GALLAIRE, H., MINKER, J. and NICOLAS, J. M. 1984. Logic and databases: A deductive approach. *Computing Surveys*, **16**, 153–185.

GHOSH, S., LIN, C. C. and SELLIS, T. 1988. Implementation of a Prolog-INGRES interface. *Sigmod Record*, **17**, 77–88.

GRAY, P. M. D. and LUCAS, R. J. 1988. *Prolog and databases*. Ellis Horwood, Chichester, 358pp.

PORTER, M. F. 1982. GOS: A package for making catalogues. *Information Technology: Research and Development*, **1**, 113–129.

—— 1983. Information Retrieval at the Sedgwick Museum. *Information Technology: Research and Development*, **2**, 169–186.

PRICE, D. 1984. Computer-based storage and retrieval of palaeontological data at the Sedgwick Museum, Cambridge, England. *Palaeontology*, **27**, 393–405.

RAWLINGS, C. J. 1988. Designing databases for molecular biology. *Nature*, **334**, 477.

ROBERTS, D. A. and LIGHT, R. B. 1980. Progress in documentation. Museum documentation. *Journal of Documentation*, **36**, 42–84.

STERLING, L. and SHAPIRO, E. 1986. *The art of Prolog: Advanced programming techniques*. MIT Press, Cambridge, Massachusetts, 427 pp.

WILLIAMS, R. 1988. The Goblin quadtree. *Computer Journal*, **31**, 358–363.

M. JENNIFER ROGERS
DESMOND T. DONOVAN

Department of Geology
University of Bristol
Bristol BS8 1RJ

MICHAEL H. ROGERS

Department of Computer Science
University of Bristol
Bristol BS8 1TR

Typescript received 23 November 1988
Revised typescript received 7 November 1989